



What I Learned From the Reproducibility Project

By Heather Kappes

If you're interested in behavioural science, you're probably familiar with the Reproducibility Project: Psychology (RP). This compilation of 100 direct replications of psychology studies, published in 2015, has garnered a lot of attention: it was named **#8 of Top 100 Stories of 2015 by Discover Magazine**, **#6 by Science News**, **#5 in Altmetric100**, **Nature Magazine's Top Science Stories of 2015**, and a **runner-up for Breakthrough of the Year by Science Magazine**, for instance.

As was intended, the project has sparked a lot of discussion. Commentators from inside and outside of psychology have reflected on what the RP implies about the trustworthiness of published findings, the best way to assess reproducibility, and how accepted practices in behavioural research should be revised. Many of the 270 authors have already weighed in on these issues in interviews or in writing. One might suspect that there is little left to be said!

However, comments about how to react to the RP and its findings have been primarily dedicated to what the field as a whole has learned and how practices in general should change. These opinions are inevitably heterogeneous, and change will take time. So, maybe there is room left to hear about lessons learned by an individual and changes that can be implemented on a much smaller scale, by someone who played a minor role in the RP.


I ran a replication of one study included in the RP (the 100 studies included in the published paper were pulled from a list of the final studies reported in articles from 3 top psychology journals in one year). Actually, the study I eventually replicated was the second one I attempted. The lead author on the first study I tried to replicate raised several objections to the study plan I generated — ^

objections I didn't think I could overcome—and so I abandoned that attempt. Her points included the fact that my participants would not be American but rather multi-national, and though they would not be students, were unlikely to have as much work experience on average as the participants she'd recruited. Her justification for these objections was sensible, but the experience was frustrating in part because the original manuscript didn't provide descriptive statistics for these sample details (making it difficult if not impossible to construct a similar sample), and also didn't make it explicit that one shouldn't expect the effect to extend outside of samples with those characteristics.

One result of that experience, for me, has been the push to think harder to think about boundary conditions or likely moderators of the effects I've identified. Whereas many students lean toward intense scepticism about generalizability—at least, that's been my impression in teaching research methods courses—researchers often tilt the other way, assuming our results say something about human nature in general, and sometimes talking that way too. It makes sense, though, to think about why that might not be the case, prior to publication, rather than only when you learn that someone is attempting to replicate your results. Explicating likely boundary conditions or moderators ahead of time could help you consider these points in a relatively even-handed way rather than one motivated by defensiveness. It should also help you feel more confident about the possibility of future replication attempts.

The second study I attempted did end up running; it was one of those considered an “unsuccessful” replication. Although in the ideal world, a scientist would probably have no vested interest in the outcome of an experiment, I was hoping the replication would succeed, for two reasons. First, considering the state of knowledge in psychology, the higher the replication rate, the better. Wouldn't we all love to work in a field that is producing and disseminating robust, reproducible results? Second, **as the critics have pointed out**, there are many ways to “fail” to find an effect. For this reason, conducting a successful replication, in the sense of finding a statistically significant effect the same as reported in the original paper, could be seen as validating one's conscientiousness and skill as a researcher. (Some critics have argued that replication researchers are **motivated to find null results**. I can only give insight into my own motives: this is not true. Moreover, as a relatively junior researcher, working on studies of those more senior, I'd need a much stronger appetite for conflict than I possess to hope to “take them down.”)

Working on the RP meant being in contact with people who are thinking carefully about how to improve our methodology. Personally, I'm rarely positioned at the leading edge of technology—have a look at my cell phone if you don't believe me! But, some of the tools such people have been developing, like the **Open Science Framework**, I've adopted and am happy to endorse. Rather than slogging through old questionnaires and data files years later, **trying to reconstruct what was done**, you spend more time at the front end, documenting the methods you'll use, analyses you'll run, and predictions you're making, and posting them to the OSF website where they can be accessed by colleagues or by the public as you choose. This documentation should make it easier for someone else to reproduce your work. Technology like the OSF would have cut out one of the time-consuming steps of the RP (“request original materials from corresponding author...follow up if no reply received...”), and in the future, hopefully answers like, “Unfortunately I don't have the video stimuli any more. I loaned out all the copies I had and didn't get them back,” will be obsolete.

These are relatively early days for replication research; we have to overcome many years of seeing such studies as uninteresting or unworthy of publication. Exactly what replication research will look like in the future is still unclear. For me, another personal lesson from working on the RP has been observing the difference between theory and practice as applied to replication. John Lynch and colleagues, **reflecting on the Replication Corner at the International Journal of Research in Marketing**, make a case for conducting conceptual (“replicating at the construct level but with different operationalization”) rather than direct replications. My experience on the RP is that the distinction between the two may not be so clear-cut. I replicated a study on differences between bilingual and monolingual individuals. The original study compared monolingual Spanish speakers in Spain to bilingual Dutch-English speakers in the Netherlands. My replication 

compared monolingual English speakers and bilingual English + second language speakers in London. Is this a direct replication, because we established the same balanced samples (e.g., matched for education level, gender, intelligence) and compared their performance on the same dependent variable task, or a conceptual replication since I operationalized “monolingual” and “bilingual” differently? Depending on how the project is framed, it may be seen as having more or less value, so these points are worth considering...but it is also worth considering that many times, the distinction may be somewhat arbitrary.

However we decide to proceed with defining best practices for replication, I would make the case that replication research is a worthwhile thing to do. One of the criticisms levied at replications in general is that they're a poor use of time, **distracting from generating and testing new ideas**. *I'm sympathetic to this argument, in part because my own inclinations run toward the surprising rather than the precise. But, most of us entered this field, rather than one where we would experience more freedom to use data to make a point rather than test an idea, because we genuinely care about advancing knowledge. The RP has not only contributed knowledge about the reproducibility of a set of findings under a set of conditions, it has supplied ideas about how to establish this reproducibility. Whether future researchers follow the RP methods or others, at least, we have a starting point. It is hard to see this as anything but an excellent use of time.*

Perhaps the key takeaway from the RP experience, for me, was the pervasive theme that it is not “us” (RP authors, or replication researchers in general) versus “them” (those who have published to-be-replicated research). Senior RP authors have repeatedly and publicly noted that the methods we are criticizing are the methods we've all been using. This can be an uncomfortable realization, but serves as a starting point for moving forward in pursuit of a better science, one individual at a time.



*Note: Heather's contribution to the Reproducibility Project was conducted in the LSE Behavioural Research Lab (BRL). As a result of this study, the lab now boasts a CRT monitor with 100Hz refresh rate, like the one pictured below, top-of-the-line technology from about 20 years ago. If you're not into monitor technology, you might be surprised to know that precision millisecond presentation of stimuli requires this kind of monitor, although you're probably not surprised to know that they are quite hard to find these days. Good news is that if you're at the LSE and looking to run a study involving very quick stimulus presentation, this monitor is waiting.

Author: Heather Kappes, Department of Management, LSE

Heather Kappes has a PhD in Social Psychology from New York University and is an Assistant Professor of Marketing in the Department of Management at the LSE. She studies self-regulation, motivation, and goal pursuit, and is currently examining these topics in the context of consumer finance.

By **Heather Kappes** | January 11th, 2016 | **Events**

Comments are closed.



